ALIGNMENT OF UNCALIBRATED IMAGES FOR MULTI-VIEW CLASSIFICATION

Sercan Ömer Arık[*]

Bilkent University
Dept. of Electrical and Electronics Engineering
TR-06800 Bilkent, Ankara, Turkey

Elif Vural[†] *and Pascal Frossard*

Ecole Polytechnique Fédérale de Lausanne
Signal Processing Laboratory (LTS4)
Lausanne, 1015 - Switzerland

ABSTRACT

Efficient solutions for the classification of multi-view images can be built on graph-based algorithms when little information is known about the scene or cameras. Such methods typically require a pairwise similarity measure between images, where a common choice is the Euclidean distance. However, the accuracy of the Euclidean distance as a similarity measure is restricted to cases where images are captured from nearby viewpoints. In settings with large transformations and viewpoint changes, alignment of images is necessary prior to distance computation. We propose a method for the registration of uncalibrated images that capture the same 3D scene or object. We model the depth map of the scene as an algebraic surface, which yields a warp model in the form of a rational function between image pairs. The warp model is computed by minimizing the registration error, where the registered image is a weighted combination of two images generated with two different warp functions estimated from feature matches and image intensity functions in order to provide robust registration. We demonstrate the flexibility of our alignment method by experimentation on several wide-baseline image pairs with arbitrary scene geometries and texture levels. Moreover, the results on multi-view image classification suggest that the proposed alignment method can be effectively used in graph-based classification algorithms for the computation of pairwise distances where it achieves significant improvements over distance computation without prior alignment.

***Index Terms*—** Image alignment, image registration, image warping, multi-view image classification, graph-based classification

1. INTRODUCTION

With the rapid development of camera arrays and vision sensor networks, multiview image classification has become an important problem. The challenge in such settings consists in taking benefit of multiple diverse observations of the same scene or objects in order to increase the performance of image analysis applications. Classification is generally achieved by comparing the observations and training data. If information is available about the scene or the camera settings, one can register images through classical computer vision techniques and then use common classification algorithms. However, there are various applications where such information is unavailable, e.g. consider the categorization of a collection of geographical or touristic images accessed through web. The classification of data through the retrieval of imaging parameters becomes less practical in this case. Graph-based methods offer effective solutions for such datasets that possess an intrinsic manifold structure [1], [2]. Such methods builds on pairwise distance computations between samples in order to approximate the geodesic distance between samples on the same manifold. This approximation however fails if the samples are captured from very different viewpoints, which brings the necessity of aligning the images before the determination of pairwise similarities.

In this work, we propose a novel method for the alignment of uncalibrated images of a scene or an object, which requires no additional information about camera parameters or scene structure. In computer vision literature, image registration is typically achieved by first retrieving the intrinsic and extrinsic camera parameters, and then obtaining a dense reconstruction of the 3D scene [3]. We rather assume that the depth map of the scene can be modeled with an algebraic representation and we write the mapping between image coordinates as a rational warp function. Since the camera parameters are implicitly included in this function, the optimization of the model parameters results in a joint camera calibration and image registration. We combine feature-based and intensity-based registration [4] in order to compute the warp function parameters as a weighted combination of two rational warp functions. It leads to robust registration for a wide range of images with varying texture levels and viewpoint changes by combining the advantages of both types of registration. The weights of the feature-based and intensity-based models in the overall model are determined by the distance of image points to feature points. The complexity of the overall method is mainly determined by the complexity of the optimization of intensity-based model parameters, which we perform by using an unconstrained simplex search method. We evaluate the performance of the proposed method experimentally and show that it provides relatively high registration accuracy with respect to some reference image registration methods. Moreover, experiments on multiview image classification show that the usage of the proposed alignment scheme before pairwise distance computation considerably improves the classification accuracy in multiview problems with semi-supervised learning and graph-based label propagation.

In image registration literature, the transformation between two images can be represented in various ways, for instance through affine and polynomial models or surface splines [5], [6]. In some intensity-based registration methods, transformation parameters are optimized based on manifold models [7], [8], [9]. Mutual information methods constitute another type of solutions for intensity-based registration, which have found many applications in the registration of magnetic resonance images [4]. Feature-based methods for the estimation of model parameters have quite favorable computation complexities, however they have the drawback that the number and

[*]The first author performed the work while at EPFL.

[†]This work has been partly supported by the Swiss National Science Foundation under grant number 200021_120060.

quality of feature matches may depend on the data. On the other hand, intensity-based registration methods usually involve complicated cost functions and may become insufficient in handling significant viewpoint changes. We rather propose to combine the advantages of both feature- and intensity-based methods in a novel hybrid and generic warp model, whose benefits are demonstrated in multi-view classification problems.

2. ALIGNMENT OF MULTI-VIEW IMAGES

Consider a setting with M different objects labeled as $\{s_1, \cdots, s_M\}$, and a set of images $\{I_i\}_{i=1}^N$, where each image captures one of these M objects. The multi-view classification problem consists in the assignment of a label s_m to each image I_i, i.e., the determination of the object present in the image.

Graph-based classification algorithms such as [1] or [10] require a pairwise distance matrix $\mathbf{D}$, where the $(i,j)^{th}$ entry of this matrix corresponds to a measure of distance between the image pair (I_i, I_j). This measure is typically taken as the Euclidean distance. However, as the amount of transformation or viewpoint change between two images of a scene becomes more significant, the validity of the Euclidean distance in the assessment of image similarity gets weaker. Thus, we concentrate on the pairwise registration of images as a preprocessing stage before the computation of the Euclidean distance.

The registration of an image pair consists in the estimation of a mapping from the coordinates of the first image to the coordinates of the second image. Given two image intensity functions $I_1(x,y), I_2(x,y) \in L^2(\mathbb{R}^2)$, we would like to compute mappings f_u and f_v, such that $I_1(x,y) = I_2(u,v)$, where $u = f_u(x,y)$ and $v = f_v(x,y)$. In the following, we assume that $I_1(x,y)$ is the reference image, $I_2(x,y)$ is the target image and both images view the same scene or object. In Sec 2.1 we derive a model for the functions f_u and f_v, where our registration algorithm will be based on the optimization of the parameters of these models such that the error between I_2 and its estimation $\hat{I}_2$ from I_1 is minimized.

For uncalibrated image pairs, the determination of exact warp functions f_u and f_v based on the 3D scene structure is difficult due to the fact that this requires the estimation of the camera parameters corresponding to both images, as well as the depth map of the scene. The computation of internal camera parameters is especially challenging when the number of available images is limited. Also, the accuracy of camera calibration may be affected by the errors in feature detection and matching. Considering that our main objective is to obtain an image similarity metric rather than 3D reconstruction, we approach this problem from a different perspective. Instead of explicitly computing the depth map, we model it as an algebraic surface and derive a novel warp model.

2.1. Geometric framework

Practical imaging systems are generally represented with the pinhole camera model. Let $p = (X,Y,Z)$ denote a 3D point in a scene captured by a stereo imaging system. We assume that the coordinate frame of the first camera is taken as the world coordinate system and the origin of the image plane is the image center. Let us denote the internal calibration matrices of the first and second cameras by $\mathbf{K}_1$ and $\mathbf{K}_2$; and the rotation and translation matrices of the stereo system by $\mathbf{R}$ and $\mathbf{t}$. We have the basic relations [11]

$$x = f_1 \frac{X}{Z}, \;\; y = f_1 \frac{Y}{Z}, \tag{1}$$

$$u = \frac{\big((f_2/f_1)\, r_{11}\, x + (f_2/f_1)\, r_{12}\, y + f_2\, r_{13}\big) Z + t_x f_2}{\big((r_{31}/f_1)\, x + (r_{32}/f_1)\, y + r_{33}\big) Z + t_z}, \tag{2}$$

$$v = \frac{\big((f_2/f_1)\, r_{21}\, x + (f_2/f_1)\, r_{22}\, y + f_2\, r_{23}\big) Z + t_y f_2}{\big((r_{31}/f_1) x + (r_{32}/f_1)\, y + r_{33}\big) Z + t_z}, \tag{3}$$

where (x,y) and (u,v) are respectively the projections of p on the first and second image planes, $\mathbf{K}_1 = diag(f_1, f_1, 1) \in \mathbb{R}^{3\times3}$, $\mathbf{K}_2 = diag(f_2, f_2, 1) \in \mathbb{R}^{3\times3}$, and $\mathbf{R}$ and $\mathbf{t}$ are given by

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}, \mathbf{t} = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix}.$$

Then we assume the following rational function model for the depth value Z of a pixel with coordinates (x,y) in the first image,

$$Z = \frac{\sum_{i=1}^{6} m_i x^{k_1} y^{r_1}}{\sum_{i=7}^{16} m_i x^{k_2} y^{r_2}}, 0 \le k_1 + r_1 \le 2, 0 \le k_2 + r_2 \le 3. \tag{4}$$

We have chosen the order of the numerator and denominator polynomials experimentally, considering the trade-off between the accuracy of the model in representing the 3D structures of typical target scenes and the complexity of model computation. The examination of camera geometry relationships (1) together with equation (4) shows that the selected model can represent many algebraic 3D surfaces such as paraboloids and planes. Combining Eq. (4) with Eqs (2) and (3), the coordinates (u,v) in the second image are obtained in terms of the coordinates (x,y) in the first image as

$$u = f_u(x,y) = \frac{\sum_{i=1}^{10} a_i x^k y^r}{\sum_{i=11}^{20} a_i x^k y^r}, 0 \le k + r \le 3, \tag{5}$$

$$v = f_v(x,y) = \frac{\sum_{i=21}^{30} a_i x^k y^r}{\sum_{i=31}^{40} a_i x^k y^r}, 0 \le k + r \le 3, \tag{6}$$

where the rotation, translation and internal camera parameters are implicitly involved in this formulation. Thus, the set of parameters $\{a_i\}$ of (5) and (6) give an approximation $\hat{I}_2(x,y)$ of the target image $I_2(x,y)$ as $\hat{I}_2(x,y) = I_1(f_u(x,y), f_v(x,y))$.

2.2. Model computation

The parameters $\{a_i\}$ of the warp model defined by Eqs (5) and (6) can be computed from a set of matched features between I_1 and I_2, as well as by using directly the image intensity functions. However, as explained in Sec. 1, both of these approaches have limitations. Therefore, in our registration framework we have chosen to build on the advantages of both approaches. We compute the overall warp model as a weighted linear combination of a feature based model $f = \{f_u, f_v\}$ and an intensity based model $g = \{g_u, g_v\}$, where f and g are both of the form given by Eqs (5) and (6) with different sets of parameters $\{a_i\}_{i=1}^{40}$ and $\{b_i\}_{i=1}^{40}$. Hence, we obtain the approximation of the target image as

$$\begin{aligned} \hat{I}_2(x,y) = {} & w(x,y)\, I_1\big(f_u(x,y), f_v(x,y)\big) \\ & + (1 - w(x,y))\, I_1\big(g_u(x,y), g_v(x,y)\big), \end{aligned} \tag{7}$$

where $w(x,y)$ is a weight map defined as a superposition of Gaussian functions of the distances between (x,y) and the feature points $\{(x^k, y^k)\}$. We compute the feature-based model f from the coordinates of a set of matched features $\{(x^k, y^k)\}$ and $\{(u^k, v^k)\}$ between the first and second images, where the model parameters

$\{a_i\}$ are given by the least-squares solution of a homogeneous equation system linear in the coordinates. In order to achieve applicability to wide-baseline images with large viewpoint changes, we use the Affine Scale Invariant Feature Transform (ASIFT) algorithm for feature detection [12], where we eliminate outliers from the set of matches using Random Sample Consensus (RANSAC) [13].

Once the feature-based warp model f and the weight map $w(x, y)$ are determined, we keep the parameters $\{a_i\}$ fixed, and optimize the parameters $\{b_i\}$ of the intensity-based model by minimizing the registration error $\|I_2(x, y) - \hat{I}_2(x, y)\|$, i.e., the norm of the difference between the target image and its estimation. We use the Nelder-Mead simplex algorithm [14] for model optimization, which is a direct simplex search method for unconstrained multidimensional optimization. We call our novel registration method Image Alignment with Parametric Modeling (IMALP).

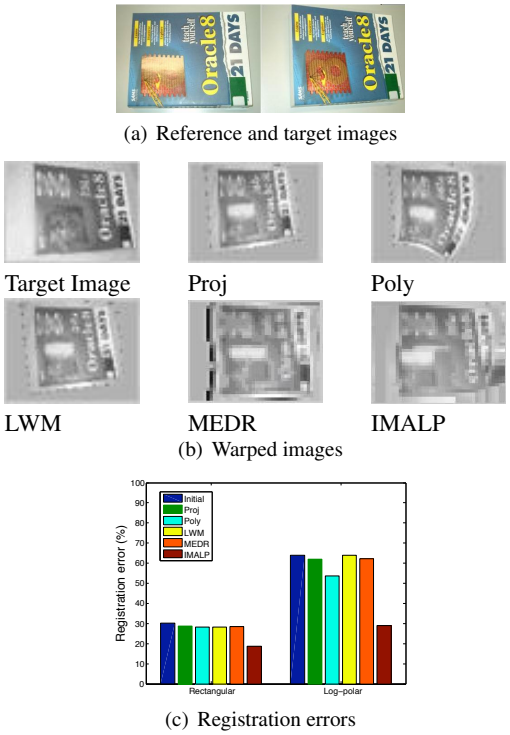
3. EXPERIMENTAL RESULTS

3.1. Experiments on Image Alignment

We first evaluate the performance of the proposed algorithm in the registration of image pairs. The accuracy of the registration is measured by the registration error E, which is defined as the norm of the difference image after warping as a percentage of the norm of the target image, i.e., $E = 100 \, \|\hat{I}_2(x, y) - I_2(x, y)\| \, / \, \|I_2(x, y)\|$.

We compare the IMALP algorithm with some state-of-the-art registration algorithms. The survey in [4] gives a review of several image registration methods. The transformation between an image pair can be modeled through global or local mappings depending on the scene structure. Once the type of the mapping is defined, model parameters can be estimated based on feature matches or image intensity functions. We compare our algorithm firstly to the basic approaches of representing the mapping between the coordinates of the two images through a polynomial model and a perspective projection model [4]. The perspective projection model assumes a planar scene structure with arbitrary normal direction. These two methods are abbreviated respectively by 'Poly' and 'Proj' in the graph legends. Another reference approach is the local weighted mean transformation model proposed in [15], denoted by 'LWM'. In this model, the warp function is a weighted sum of different polynomial functions, where the weights vary locally. In all these three methods ('Poly', 'Proj' and 'LWM') model parameters are estimated from feature matches. Finally, we test also the medical image registration method in [6] (abbreviated as 'MEDR'), which uses a global affine model with smoothly varying local parameters. The original unregistered image, which corresponds to $\hat{I}_2(x, y) = I_1(x, y)$, is shown as 'Initial'. The methods 'Poly' and 'Proj' have quite low computational complexities as the model parameters are computed only from feature matches. Even though 'LWM' is a feature-based method, it has a slightly increased complexity compared to 'Poly' and 'Proj' as it additionally involves the computation of local weights. Since 'MEDR' and IMALP perform area-based optimization, they are computationally more complex than purely feature-based methods. We have observed that these two methods have similar typical runtimes. The complexity of our method is mainly determined by the complexity of the area-based model computation, which depends on the image resolution (linearly) and the Nelder-Mead simplex algorithm [14].

We perform experiments on some images selected from the ETH object images database[1]. All image pairs are chosen randomly with arbitrary camera orientations. Minor artifacts due to shading and background differences are not preprocessed and images are downsampled. For the book images shown in Figure 1(a), the warped images obtained by the tested registration methods are displayed in Figure 1(b). The registration errors obtained in the original rectangular image domain and the log-polar image domain are plotted in Figure 1(c). The results indicate that feature based algorithms show better performance for high-textured data that yield a sufficient number of matches, whereas intensity based algorithms are more preferable for less textured data as expected. As an efficient combination of both approaches, IMALP algorithm provides considerably accurate registration results in all cases. The benefits of IMALP has also been observed in the alignment of face images from FacePix database[2]. Due to the algebraic surface model adopted, possible target application areas of this method can be the registration of images with simple and smooth depth fields or the local analysis of a scene, rather than the global registration of a scene with a complicated depth field with discontinuities.



(a) Reference and target images

(b) Warped images

(c) Registration errors

Fig. 1. Registration results obtained on book images

3.2. Experiments on Multi-View Classification

Now we demonstrate the benefit of the proposed registration algorithm in multi-view classification applications. Given a set of N multi-view images $\{I_i\}_{i=1}^{N}$ of M different scenes, for each image pair (I_i, I_j) in the set we warp the reference image I_i to obtain an approximation $\hat{I}_j^i$ of the target image I_j with the IMALP algorithm. Then we construct a symmetric distance matrix $\mathbf{D}$, where the $(i, j)^{th}$ entry of the matrix is the total registration error

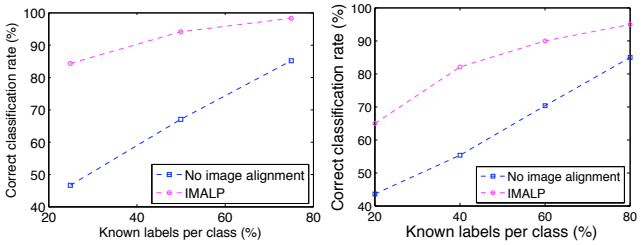
$$D_{ij} = \frac{\|\hat{I}_j^i(x, y) - I_j(x, y)\|}{\|I_j(x, y)\|} + \frac{\|\hat{I}_i^j(x, y) - I_i(x, y)\|}{\|I_i(x, y)\|}. \quad (8)$$

[1] http://www.vision.ee.ethz.ch/datasets/index.en.html

[2] http://www.facepix.org

We then perform experiments where we use the label propagation algorithm for semi-supervised classification. Label propagation [10] is one of the popular algorithms in semi-supervised learning that applies to sets of data whose class labels are partially known [2]. The algorithm is based on smoothness assumptions of the data and requires a similarity matrix and a partially known label matrix. The similarity matrix $\mathbf{S}$ is obtained from the distance matrix $\mathbf{D}$ as $\mathbf{S} = \mathbf{G}^{-0.5}\mathbf{W}\mathbf{G}^{-0.5}$, where $\mathbf{W}$ is the weight matrix defined as $W_{ij} = \exp(\frac{-D_{ij}}{\sigma^2})$, and $\mathbf{G}$ is a diagonal matrix given by $G_{ii} = \sum_{j=1}^{n} W_{ij}$. Based on the similarity matrix, one can perform the classification of the multiple observations following [1].

We use a set of images from the ETH objects database without any preprocessing except downsampling. We compare the classification accuracy obtained with the prior alignment of images using IMALP algorithm with respect to the classification accuracy obtained with no image alignment (in this second case, we compute the classification error by computing a $\mathbf{D}$ matrix from Eq. (8) without registration). We first experiment on a data set consisting of 4 different images of 7 objects captured from random perspectives. Each object is considered as a separate class. We repeat the experiment 50 times, where we assign the known class labels among the images of the same object randomly at each run. We plot the correct classification rate with respect to the ratio of known labels per class in Figure 2(a). The results are averaged over all runs. Then we repeat the same experiment on the face image data set consisting of 5 images of 4 subjects from FacePix database. The camera angle variation between consecutive images of the same subject is 30°. The classification rates are plotted in Figure 2(b), again averaged over 50 runs with randomly assigned labels. The results show that the classification performance of the label propagation algorithm is improved considerably when the proposed registration algorithm is used for the pairwise alignment of images before distance computation. We note that the classification accuracy obtained with such a graph-based algorithm is expected to be highly affected by the accuracy of the registration method used for prior alignment.



(a) Object images (b) Face images

Fig. 2. Semi-supervised classification with prior image alignment

4. CONCLUSION

We have proposed a registration method for the alignment of uncalibrated multi-view images of objects, which requires no prior information about the capturing system or scene. We model the depth map of the scene as a rational function and obtain a parametric warp model between the reference and target images. The aligned version of the reference image with respect to the target image is a weighted combination of two images, which are yielded by two different warp functions computed from feature matches and image intensities. We show that the proposed algorithm can be applied flexibly under large viewpoint changes and for different scene structures and that it outperforms state-of-the-art image registration methods. We further demonstrate its benefits in the classification of multi-view images using graph-based algorithms where preprocessing the images with the described registration method contributes significantly to the accuracy of classification.

5. REFERENCES

[1] E. Kokiopoulou and P. Frossard, "Graph-based classification of multiple observation sets," *Pattern Recognition*, July 2010.

[2] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning", *Synthesis Lectures on Artificial Intelligence and Machine Learning 6*, Morgan and Claypool Publishers, 2009.

[3] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch, "Visual modeling with a hand-held camera", *International Journal of Computer Vision*, vol. 59, no.3, pp. 207-232, 2004.

[4] B. Zitova, "Image registration methods: a survey," *Image and Vision Computing*, vol. 21, no. 11, pp. 977-1000, October 2003.

[5] L. Zagorchev and A. Goshtasby, "A comparative study of transformation functions for nonrigid image registration," *IEEE Transactions on Image Processing*, vol. 15, no. 3, pp. 529-538, February 2006.

[6] S. Periaswamy and H. Farid, "Medical image registration with partial data," *Medical Image Analysis*, vol. 10, pp. 452-464, 2006.

[7] A. Fitzgibbon and A. Zisserman, "Joint manifold distance: A new approach to appearance based clustering," *IEEE Computer Society Conference on Computer Vision and Patten Recognition*, 2003.

[8] E. Kokiopoulou and P. Frossard, "Minimum distance between pattern transformation manifolds: Algorithms and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no.7, pp. 1225-1238, 2009.

[9] N. Vasconcelos and A. Lippman, "A multiresolution manifold distance for invariant image similarity," *IEEE Transactions on Multimedia*, vol. 7, no.1, pp. 127-142, 2005.

[10] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," *Technical report CMU-CALD-02-107*, Carnegie Mellon University, Pittsburgh, 2002.

[11] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, April 2004.

[12] J. M. Morel and G. Yu, "ASIFT: A New Framework for Fully Affine Invariant Image Comparison," *SIAM Journal on Imaging Sciences*, vol. 2 no. 2, pp. 438-469 ,2009.

[13] P. Torr and D. Murray, "The development and comparison of robust methods for estimating the fundamental matrix," *International Journal of Computer Vision*, vol. 24, pp. 271-300, 1997.

[14] J. A. Nelder and R. Mead, "A Simplex Method for Function Minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308-313, 1965.

[15] A. Goshtasby, "Image registration by local approximation methods," *Image and Vision Computing*, vol. 6, pp. 255-261, 1988.